WAGENINGEN
UNIVERSITY & RESEARCH

# Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining.

Jasper J. Koehorst[1,*] Jesse C.J. van Dam[1,*], Jon Olav Vik[2], Peter J. Schaap[1] and Maria Suarez-Diez[1]

1. Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708 WE, Wageningen, The Netherlands

2Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences (IHA), Faculty of Life Sciences (BIOVIT), Norwegian University of Life Sciences (NMBU),

*Authors contributed equally

## Background

A standard structured format is used by public (bio) sequence databases to present genome annotations. The current format provides limited support for data provenance and data mining, hampering comparative analyses at large scale.

We have developed **Genome Biology Ontology Language** (**GBOL**) and associated infrastructure (**GBOL stack**). GBOL provides a consistent representation of functional genome annotations linked to the provenance.

## GBOL ontology

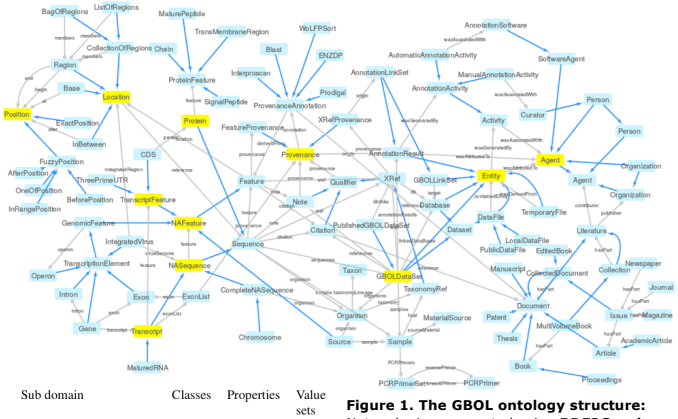

**Figure 1. The GBOL ontology structure:** Network view generated using **RDF2Graph.**

| Sub domain | Classes | Properties | Value sets |
|---|---|---|---|
| Genomic locations | 16 | 17 | 1 |
| Genes transcripts and features | 114 | 133 | 17 |
| Document structure | 27 | 107 | 7 |
| Dataset-wise provenance | 22 | 54 | 0 |
| Element-wise provenance | 5 | 9 | 0 |
| BIBO | 59 | 90 | 2 |

**Table** 1: key characteristics. Of the GBOL ontology

## Links to existing ontologies

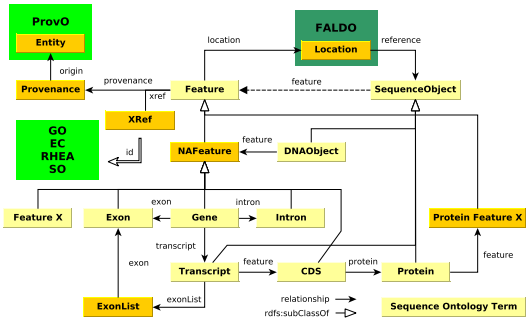


**Figure 2.** Embedding of the GBOL ontology with already existing ontologies such as FALDO, ProvO, GO, EC, RHEA and SO.

## An interoperable genome annotation pipeline

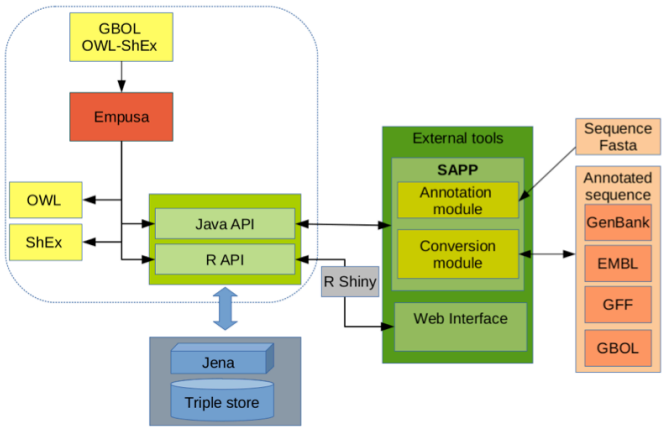


**Figure 3**: The **GBOL stack** (dashed box) provides **GBOL** (Yellow) and an infrastructure to keep it consistent and extendable based on **Empusa**. **SAPP** retrieves raw genome data from the triple store, runs genome annotation and uses the GBOL ontology to store predictions and data provenance directly as RDF triples (Blue).
Functional annotations, data provenance and linked meta-data can be queried within JAVA and R with SPARQL and by using a web interface (Green). Parsers have been developed for conversion of annotation files in standard formats (Orange).

## The Empusa code generator.

**Empusa** can be used to define an ontology and generate an associated application programming interface (API) that can be used to perform data consistency checks.
The use of Empusa ensures consistency within and between the ontology (OWL), the Shape Expressions (ShEx) describing the graph structure and the content of the resource.

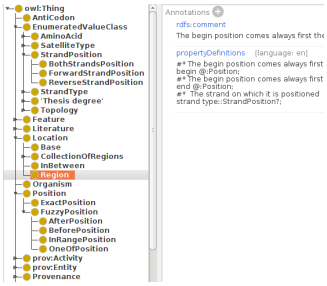

**Figure 4. Example Empusa definition format.** Properties within a class can be defined with the propertyDefinitons annotation property.

## Availability

- **RDF2graph**:van Dam et al. *Journal of Biomedical Semantics* 2015
  - https://github.com/jessevdam/RDF2Graph
- **Empusa**: bioRxiv
  - https://gitlab.com/Empusa
- **GBOL:**
  - https://gitlab.com/gbol/ Documentation & namespace: https://gbol.life/0.1/
- **SAPP** Koehorst et al *Bioinformatics* 2018
  - https://gitlab.com/sapp

This work has received funding from the Research Council of Norway, No. 248792 (DigiSal) and from the European Union FP7 and H2020 under grant agreements No. 305340 (INFECT), No. 635536 (EmPowerPutida) and No. 634940 (MycoSynVac).

INFECT EU-FP7-HEALTH
MYCOSYNVAC
EMPOWER PUTIDA
1000100 1010011 DIGITAL SALMON

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining
Jesse C.J. van Dam, Jasper J. Koehorst, Jon Olav Vik, Peter J. Schaap, Maria Suarez-Diez
doi: https://doi.org/10.1101/184747